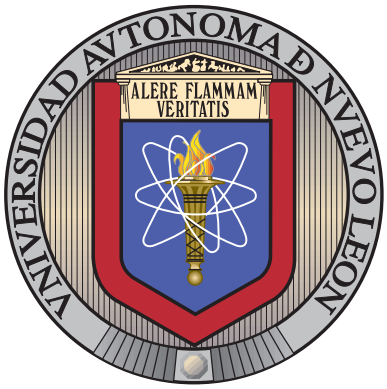
Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Subdirección Académica



Modelado y visualización de relaciones entre contaminantes del aire y salud pública

por

Selene Berenice Prado Prado

como requisito parcial para obtener el grado de

ingeniería en tecnología de software

Julio 2022

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Subdirección Académica

Modelado y visualización de relaciones entre contaminantes del aire y salud pública

por

Selene Berenice Prado Prado

como requisito parcial para obtener el grado de

ingeniería en tecnología de software

Julio 2022

Universidad Autónoma de Nuevo León

Facultad de Ingeniería Mecánica y Eléctrica

Subdirección Académica

Los miembros del Comité de Tesis recomendamos que la Tesis «Modelado y visualización de relaciones entre contaminantes del aire y salud pública», realizada por el alumno Selene Berenice Prado Prado, con número de matrícula 1810042, sea aceptada para su defensa como requisito parcial para obtener el grado de Ingeniería en Tecnología de Software.

El Comité de Tesis

Dra. Satu Elisa Schaeffer

Asesora

Dra. Sara Elena Garza Villarreal

Coasesora

Dr. José Arturo Berrones Santos

Revisor

Vo. Bo.

Dr. Fernando Banda Muñoz

Subdirección Académica

San Nicolás de los Garza, Nuevo León, julio 2022

Índice general

Índice de figuras

Índice de cuadros

Agradecimientos

Quiero agradecer a la Dra. Elisa por el apoyo durante el desarrollo de mi tesis y por la motivación y conocimientos brindados para seguir desarrollandome profesionalmente en lo que me gusta. Al programa PAICYT-UANL por su contribución brindada bajo las claves CE1421-20 y CE1842-21.

A mis padres, Lilia Prado López y Adan Alfaro Lerma, por su apoyo y motivación constante desde siempre. A mis hermanos Angel, Estrella, y Adali, a quienes he visto crecer y de quienes he aprendido mucho.

Resumen

Selene Berenice Prado Prado.

Candidato para obtener el grado de Ingeniería en Tecnología de Software.

Universidad Autónoma de Nuevo León.

Facultad de Ingeniería Mecánica y Eléctrica.

Título del estudio: Modelado y visualización de relaciones entre contaminantes del aire y salud pública.

Número de páginas: 34.

Objetivos y método de estudio: El objetivo de la investigación es generar modelos que permitan visualizar relaciones entre contaminantes atmosféricos y salud pública. Los modelos generados se utilizan en conjunto con datos obtenidos de la Secretaría de Salud del Gobierno de México y registros de los niveles de los contaminantes presentes en el área metropolitana de Monterrey.

El tener un modelo que permita visualizar relaciones entre contaminantes atmosféricos y salud pública que sea utilizado con datos confiables y verídicos pueden ayudar a visualizar el impacto que tiene el aumento del nivel de contaminantes atmosféricos.

Contribuciones y conclusiones: Durante la investigación...

Firma de la asesora: _____

Dra. Satu Elisa Schaeffer

Firma de la coasesora: _____

Dra. Sara Elena Garza Villarreal

Capítulo 1

Introducción

El crear modelos para la visualización de datos ayuda a observar con mayor claridad los datos para encontrar relaciones entre ellos.

El *aprendizaje máquina*[1] es un área dentro de la *ciencia de datos*[2] que puede ayudar a crear dichos modelos para tener una más eficiente visualización cuando se trabaja con una gran cantidad de datos, que es lo que se requiere para el presente trabajo. El área de la ciencia de datos es muy útil ya que permite trabajar con grandes cantidades de datos aminorando la cantidad de tiempo empleado en la creación de gráficos que permitan visualizar los datos.

La tarea en el presente proyecto es utilizar modelos para visualizar las relaciones entre los contaminantes del aire y salud pública, para ello se requieren datos sobre salud pública y sobre los niveles de contaminantes del aire.

Para la realización de los experimentos se tienen datos de ingresos hospitalarios provenientes de la base de datos de la Secretaría de Salud del Gobierno de México [9]. También se tienen registros de los niveles de algunos contaminantes del aire presentes en el área metropolitana de Monterrey, dichos registros son hechos por las

[1]Traducido como *machine learning* en inglés, tiene como objetivo desarrollar técnicas que les permitan a las computadoras aprender.

[2]Traducido como *data science* en inglés, involucra métodos para extraer conocimiento de datos, eso con la finalidad de que haya un mejor entendimiento de los datos.

estaciones de monitoreo pertenecientes al Sistema Integral de Monitoreo Ambiental (SIMA) [20] mostradas en la figura 1.1.

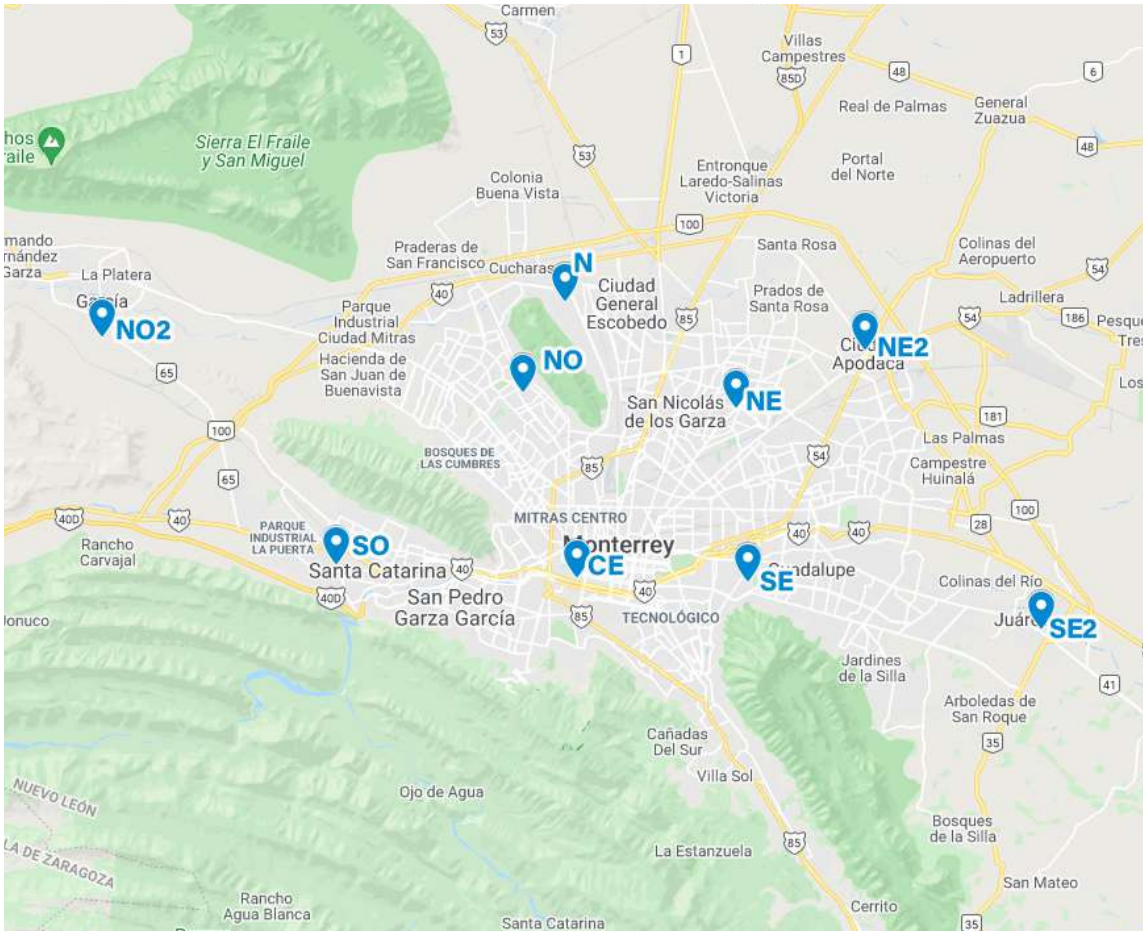


Figura 1.1: Localización de las estaciones de monitoreo de la calidad del aire.

1.1 Motivación

Existen investigaciones que ya han estudiado las relaciones entre contaminantes del aire y salud pública, sin embargo, con el presente trabajo se busca aportar a la creación de nuevas herramientas que permitan observar y estudiar dichas relaciones. El poder visualizar dichas relaciones puede ayudar a tomar medidas adecuadas que permitan aminorar los efectos negativos de los contaminantes del aire en la salud.

1.2 Hipótesis

Se plantea que con modelos de regresión se pueden obtener gráficos donde se pueden observar las relaciones entre el número de ingresos hospitalarios y los niveles de contaminantes del aire.

1.3 Objetivos

En esta sección se establece el objetivo general y los objetivos específicos sobre los que se orienta el presente trabajo.

1.3.1 Objetivo general

El objetivo de generar, implementar y evaluar modelos que muestran las relaciones existentes entre contaminantes del aire y salud pública tiene la finalidad de apoyar a la implementación de estrategias que aminoran los efectos negativos de los contaminantes del aire en la salud de las personas. Con los modelos generados se puede tener una herramienta que permite identificar gráficamente las relaciones con solo proporcionarle el conjunto de datos.

1.3.2 Objetivos específicos

- Generar, implementar y evaluar modelos de regresión que permite cuantificar las relaciones entre contaminantes del aire y salud pública a partir de un conjunto de datos.

- Diseñar e implementar visualizaciones interactivas que permiten explorar los modelos implementados y su validez estadística.

- Evaluar la eficacia de los modelos generados para que así al utilizar cualquiera de los modelos generados se pueda tener noción de la fiabilidad del análisis realizado a partir de los resultados producidos por los modelos.

Capítulo 2

Antecedentes

Existen factores ambientales que afectan la salud de una comunidad como: el abastecimiento de agua potable y el saneamiento, la vivienda y el hábitat, la alimentación, la contaminación ambiental, el empleo de productos químicos y los riesgos ocupacionales [6].

Contaminación del aire es un término usado para describir la presencia de uno o más contaminantes en la atmósfera, cuyas cantidades y características pueden resultar perjudiciales o interferir con la salud, el bienestar u otros procesos ambientales naturales [7].

En el presente capítulo se presentan los fundamentos y definiciones de los conceptos más relevantes para el tema de estudio abordado.

2.1 Monitoreo de calidad del aire

Existen diversos estudios que muestran que existen potenciales efectos a la salud cuando en el aire están presentes contaminantes en forma de partículas, gases o agentes biológicos.

Korc y Sáenz [15] mencionan que desde inicios de 1950 se observa una preocupación por los contaminantes del aire en los países de América Latina y el Caribe. Las universidades y dependencias de los ministerios de salud fueron los organismos que realizaron las primeras mediciones de contaminación en el aire.

En 1965, el Consejo Directivo de la Organización Panamericana de la Salud (OPS) recomendó el establecer programas de investigación de la contaminación del agua y del aire, con el objetivo de colaborar en el desarrollo de políticas adecuadas de control [13].

Mediante el Centro Panamericano de Ingeniería Sanitaria y Ciencias del Ambiente (CEPIS), la OPS acordó establecer una red de estaciones de muestreo de la contaminación del aire. En junio de 1967 La Red Panamericana de Muestreo Normalizado de la Contaminación del Aire (REDPANAIRE) inició sus operaciones recolectando muestras mensuales de polvo sedimentable (PS) y muestras diarias de partículas totales en suspensión (PTS) y de SO2. La REDPANAIRE comenzó con ocho estaciones y a fines de 1973 tenía un total de 88 estaciones distribuidas en 26 ciudades de 14 países [13].

Para diciembre de 1973 se habían recolectado más de 350,000 datos sobre la calidad del aire, en los que se observa que algunas ciudades mostraban una tendencia al incremento de los niveles de contaminación [13].

En 1980 la REDPANAIRE desapareció y pasó a formar parte del Programa Global de Monitoreo de la Calidad del Aire, iniciado en 1976 por la OMS y el Programa de las Naciones Unidas para el Medio Ambiente (PNUMA), como parte de un sistema global de monitoreo ambiental llamado GEMS por sus siglas en inglés *Global Environmental Monitoring System*.

En la década de 1990, la OMS organizó, con carácter global, el Sistema de Información para el Control de la Calidad del Aire llamado AMIS por sus siglas en inglés *Air Management Information System*. Entre las actividades más destacadas de AMIS se incluye el coordinar las bases de datos sobre temas relacionados con la

calidad del aire.

En Nuevo León, México, las operaciones de la Red Automática de Monitoreo Atmosférico iniciaron en 1993. Dicha red en sus inicios contaba con cinco estaciones fijas de monitoreo continuo de monóxido de carbono (CO), dióxido de azufre (SO2), óxidos de nitrógeno (NOx), ozono y PM10 [15]. Como se muestra en la figura 1.1, actualmente se cuenta con nueve estaciones fijas.

2.2 Series de tiempo

Korc y Sáenz [15] mencionan que las relaciones entre niveles de concentraciones de contaminantes del aire y los efectos sobre la salud generalmente son obtenidas de estudios epidemiológicos de series de tiempo. Uno de los diseños epidemiológicos más utilizados son los estudios de series temporales. Con esos diseños se analizan las variaciones en el tiempo de la exposición al contaminante y el indicador de salud estudiado en una población [2].

Las series de tiempo se pueden definir como un conjunto de observaciones ot tomadas en un tiempo t determinado. Los estudios de series de tiempo relacionan estadísticamente los cambios temporales en la repercusión de cambios en la concentración de un contaminante en la población [5].

Para mostrar datos en una serie de tiempo, especialmente en el área médica, estos suelen agruparse en *semanas epidemiológicas*[1].

[1]Una semana epidemiológica es un estándar de medición temporal que se utiliza para comparar datos en ventanas de tiempo definidas. La primera semana epidemiológica del año termina el primer sábado de enero de cada año [1].

2.3 Clasificación de enfermedades

Existe un instrumento estadíistico y sanitario para identificar enfermedades llamado Clasificación Internacional de Enfermedades (CIE), cuya finalidad es entender las causas de morbilidad y mortalidad de la población y así mejorar la calidad de vida de la misma. Es en base a un criterio epidemiológico y sanitario establecido por Farr a finales del siglo XIX que esta clasificación agrupa enfermedades en epidémicas, generales, locales ordenadas por origen geográfico, trastornos del desarrollo y lesiones [18]. Para lograr distinguirlas se emplea un código alfanumérico que consiste de una letra en la primera posición, seguida de dos dígitos, un punto decimal y un último dígito. El rango de valores va de A00.0 a Z99.9.

2.4 Modelos de regresión lineal

La tendencia w_0 de una serie de tiempo puede ser obtenida a partir de una regresión lineal de la misma [8]. Una regresión lineal es una metodología inferencial supervisada que busca predecir valores y dado un vector de variables de entrada t por medio del ajuste de coeficientes w de la función lineal

$$\hat{y}(t, w) = w_0 + w_1 x_1 + ... + w_t x_t. \tag{2.1}$$

2.4.1 Regresión múltiple

Un modelo de regresión múltiple es un modelo complemento de la regresión lineal simple, el cual tiene dos o más variables independientes k que pueden influir en una variable dependiente y. Peniche-Camps y Cortez-Huerta [19] expresan la regresión múltiple mediante la siguiente ecuación:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon. \tag{2.2}$$

2.5 Modelos ARIMA

Los modelos autorregresivos integrados de media móvil o ARIMA, por su abreviatura en inglés abarcan un catalogo de aproximaciones para el estudio de series de tiempo.

Los modelos ARIMA utilizan variaciones y regresiones de datos estadísticos con el fin de encontrar patrones para una predicción hacia el futuro.

Capítulo 3

Estado del arte

En el presente capítulo se estudia literatura reciente relacionada con el presente trabajo, esto con el objetivo de revisar distintos métodos para resolver el problema planteado en el presente trabajo y, además, también revisar implementaciones similares para resolver problemas distintos. Lo anterior tiene la finalidad de comparar los trabajos revisados e identificar áreas de oportunidad en ellos.

En la primera sección, *trabajos relacionados*, se recopilan obras con características relacionadas al presente trabajo, ya sean relacionados con el problema que se pretende resolver o con los métodos empleados para buscar su resolución.

En la segunda sección, *análisis comparativo*, se comparan las distintas características de los trabajos revisados, de esa forma se pueden determinar las principales ventajas y desventajas de cada trabajo.

Finalmente, en la tercera sección, *áreas de oportunidad*, se realiza una conclusión acerca de los resultados obtenidos del análisis comparativo.

3.1 Trabajos relacionados

Se recopila literatura relacionada desde el año 2017 hasta el año 2021. En esta sección los trabajos se mencionan en orden cronológico tomando en cuenta su año de publicación.

Martín y Bayle [17] estudian la relación entre los niveles de contaminantes ambientales y la presencia de casos de enfermedades respiratorias en las consultas pediátricas. La variable dependiente analizada es la demanda en las consultas pediátricas por bronquiolitis, episodios de broncoespasmo y procesos respiratorios de vías altas. Como variables independientes se tienen los valores de contaminación ambiental. Se calculan coeficientes de correlación y regresión lineal múltiple.

Guarnaccia *et al.* [11] abordan la necesidad de monitoreo, control y predicción de la pendiente de los niveles de contaminantes del aire. Para abordar el problema de investigación utilizan modelos ARIMA.

FC Lichtenfels *et al.* [10] estudian la asociación entre la exposición a largo plazo a la contaminación del aire y la metilación del ADN. Para ello realizan un estudio utilizando modelos de regresión lineal robustos para analizar la asociación entre la exposición al NO2 y a las partículas PM10 y PM2.5.

Zhang *et al.* [22] en su estudio abordan los niveles de contaminación del aire y su asociación con la presencia de presión sanguínea elevada en niños y adolescentes. La exposición a partículas PM10 y PM2.5 son estimadas con un modelo espacio-temporal. Son utilizados modelos lineales de efectos mixtos y modelos de regresión logística para investigar la asociación entre la exposición a partículas PM y presión sanguínea e hipertensión.

Kim *et al.* [14] estudian la relación entre los niveles de contaminación del aire y la obesidad y problemas cardiometabolicos. Para dicho estudio emplean modelos de regresión lineal.

Liu *et al.* [16] examinan las asociaciones entre la exposición temprana a la contaminación del aire y la incidencia de asma y rinitis alérgica desde el nacimiento hasta la adolescencia. Para su estudio utilizan modelos de regresión.

To *et al.* [21] estudian la asociación entre la exposición temprana a los contaminantes del aire y los egresos hospitalarios por asma. Para su estudio aplican modelos de regresión logística para el análisis de datos.

Breton *et al.* [4] abordan el estudio de la relación entre los niveles de contaminación del aire y el número de admisiones hospitalarias. Para ello se construye un modelo basado en la distribución de Poisson.

Gupta *et al.* [12] estudian la relación entre la mortalidad del coronavirus (COVID-19) y la contaminación del aire. Para dicho estudio emplean un modelo de regresión lineal para establecer la relación entre los parámetros de la contaminación del aire (concentraciones de PM10 o PM2.5) y la variable de respuesta (porcentaje mortalidad por unidad de casos reportados).

3.2 Comparación de trabajos

La mayoría de los trabajos encontrados emplean modelos de regresión lineal o modelos de predicción. Además, en todos los trabajos encontrados el problema tratado presenta una alta relación con el problema abordado en el presente trabajo de tesis. El análisis comparativo de los trabajos relacionados se hace en base de los siguientes puntos:

Modelos de regresión lineal: Son aquellos que ayudan a estudiar la relación entre una variable dependiente y una o más variables independientes.

Modelos de predicción: Son aquellos que ayudan a hacer predicciones de una variable.

Evaluación de modelos: Se refiere a la utilización de técnicas para evaluar la eficacia de los modelos generados.

Estudio de contaminantes del aire: Se refiere a que el tema de estudio incluya uno o más contaminantes del aire.

Estudio de problemas de salud: Se refiere a que el tema de estudio incluya uno o más problemas de salud.

En el cuadro 3.1 se desglosan que características presentes que se pueden encontrar en las investigaciones citadas y su relación con la investigación con la que se está trabajando actualmente.

Cuadro 3.1: Comparación de trabajos frente al desarrollado, donde ✓ indica que cumple con esta característica y × no cumple con esta característica.

Trabajo	Modelos de regresión lineal	Modelos de predicción	Evaluación de modelos	Estudio de contaminantes del aire	Estudio de problemas de salud
Martín y Bayle [17]	✓	×	×	✓	✓
Guarnaccia *et al.* [11]	×	✓	✓	✓	×
FC Lichtenfels *et al.* [10]	✓	✓	×	✓	✓
Zhang *et al.* [22]	✓	✓	×	✓	✓
Kim *et al.* [14]	✓	×	×	✓	✓
Liu *et al.* [16]	×	✓	✓	✓	✓
To *et al.* [21]	×	×	×	✓	✓
Breton *et al.* [4]	✓	✓	×	✓	✓
Gupta *et al.* [12]	✓	✓	×	✓	✓
El presente trabajo	✓	✓	✓	✓	✓

3.2.1 Áreas de oportunidad

Como se puede observar en el cuadro 3.1, la mayoría de los trabajos encontrados abordan el estudio de los contaminantes del aire y salud con excepción de Guarnaccia *et al.* [11] que se enfocan en la predicción de niveles de contaminantes del aire, lo cual puede indicar que la relación entre los contaminantes del aire y salud es un tema de relevancia en la actualidad.

Ya que la mayoría de los trabajos encontrados estudian la relación entre contaminantes del aire y salud, la mayoría de los trabajos emplean modelos de regresión lineal por que es una buena opción para el estudio de relaciones entre variables. Las excepciones, además de la ya anteriormente mencionada, son Liu *et al.* [16] y To *et al.* [21] quienes emplean otros tipos de modelos de regresión.

En el presente trabajo se elaboran modelos de predicción para el tratamiento de los datos empleados para los experimentos, ya que como menciona Zhang *et al.* [22], una de las limitaciones en este tipo de estudios es los campos sin llenar en los registros de datos.

En el presente trabajo también se emplean técnicas para evaluar los modelos generados. Solo en tres de los trabajos encontrados se aborda la evaluación de los modelos empleados, y al ser incluida en el presente estudio, puede representar una distinción.

Capítulo 4

Solución propuesta

En el presente capítulo se presenta la propuesta de diseño de la solución para el problema de investigación abordado en el presente trabajo, así como su implementación.

4.1 Diseño de la solución propuesta

En diseño de la solución propuesta se plantean las herramientas utilizadas y los pasos seguidos para la solución propuestas.

Las herramientas utilizadas en la presente investigación se muestran en el cuadro 4.1.

Cuadro 4.1: Herramientas utilizadas.

Herramienta	Versión	URL
`Python`	3.8.8	`https://www.python.org/`
`NumPy`	1.20.1	`http://www.numpy.org/`
`Pandas`	1.2.4	`https://pandas.pydata.org/`
`Seaborn`	0.11.1	`https://seaborn.pydata.org/`
`Matplotlib`	3.3.4	`https://matplotlib.org/`
`Statsmodels`	0.12.2	`https://www.statsmodels.org/`
`Scikit-learn`	0.24.1	`https://scikit-learn.org/`

4.1.1 Recolección de datos

La primera fase es la recolección de datos. El objetivo es tener un archivo que contenga datos de los niveles de uno o más contaminantes del aire en años recientes y también del mismo lugar tener datos del número de egresos hospitalarios durante esos años.

4.1.2 Selección y agrupación de datos

Después de la recolección de datos se procede a seleccionar que datos van a ser utilizados para los experimentos. Para ello se utiliza `Python` con la librería `Pandas` que permite la manipulación de datos 4.1. Para la selección y agrupación de datos se sigue el procedimiento mostrado en la figura 1.

4.1.3 Visualización de la evolución de las variables

Al ya tener seleccionados los datos a utilizar se procede a elaborar gráficos en `Python` 4.1 que muestran la evolución de las variables en el tiempo. Para ello se

Algoritmo 1 Selección y agrupamiento de datos

1: $a \leftarrow$ años de los que se obtuvieron datos
2: **for** $i \in a$ **do**
3: *contaminantes* $\leftarrow$ nombre del archivo .csv que contiene los datos de los contaminantes en el año i
4: Leer en *contaminantes* las columnas fecha y contaminante
5: *egresos* $\leftarrow$ nombre del archivo .csv que contiene los datos de los contaminantes en el año i
6: Leer en *contaminantes* las columnas fecha, padecimiento y estado
7: *estado* $\leftarrow$ estado del que se quieren obtener datos
8: Seleccionar en *contaminantes* los datos del *estado*
9: **end for**

generan los tipos de gráficos discutidos a continuación.

4.1.3.1 Series de tiempo

Se realizan series de tiempo en Python con ayuda de la librería `Matplotlib`, `Scikit-learn` y `Seaborn`, ya que son herramientas accesibles que ayudan a la generación de este tipo de gráficos 4.1. En la figura 4.1 se muestra un ejemplo de las series de tiempo generadas.

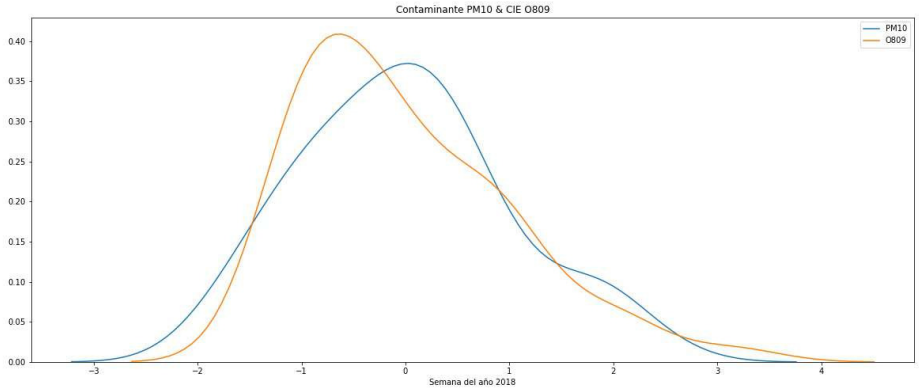


Figura 4.1: Evolución de los niveles de PM10 y el número de egresos diagnosticados la CIE O809 en el 2018.

4.1.3.2 Gráficos de radar

Los gráficos de radar o diagramas de telaraña son otra manera de visualizar un conjunto de datos. Sirven para comparar variables visualizando si existen valores o patrones de evolución en el tiempo similares entre ellas. Es por ello que en el presente trabajo se elaboran gráficos de radar con ayuda de `Python` y las librerías `NumPy` y `Matplotlib` 4.1. En la figura 4.2 se muestra un ejemplo de los gráficos de telaraña generados.

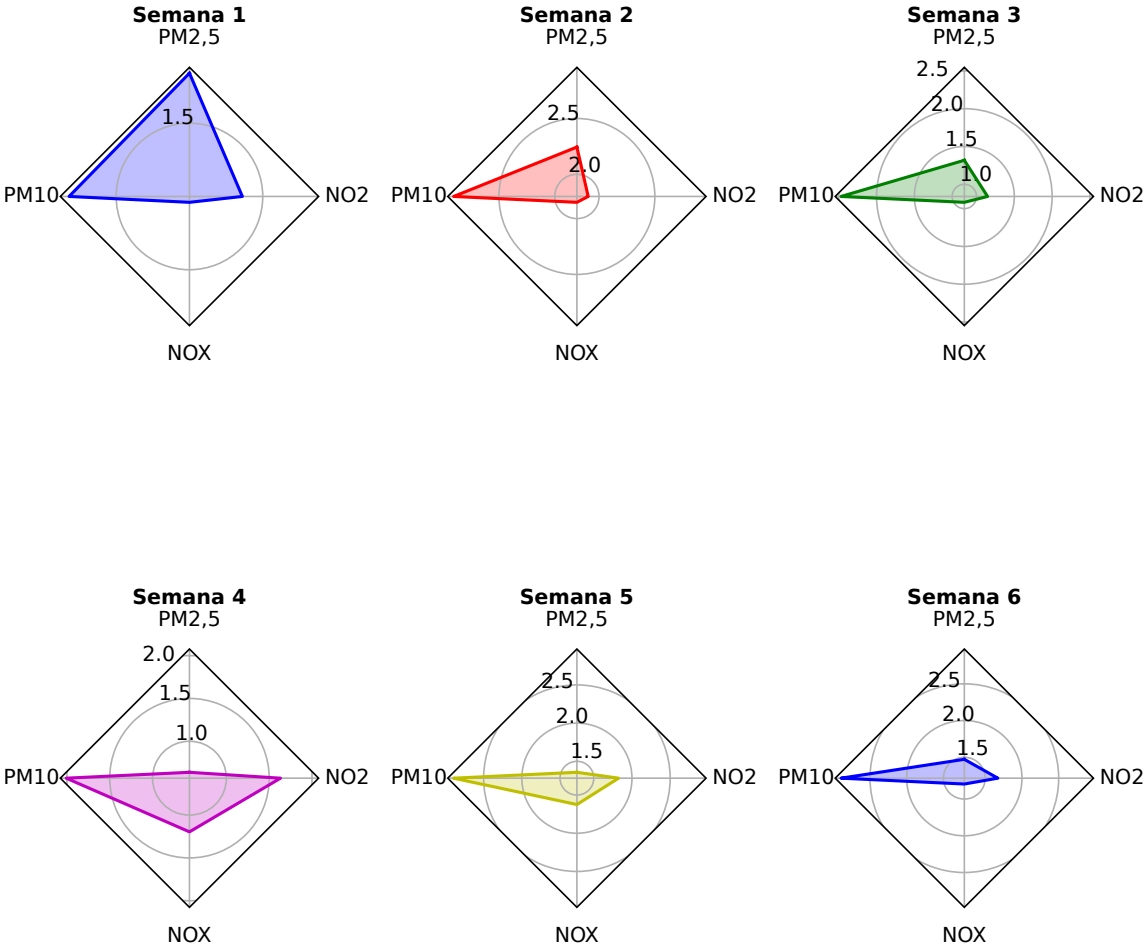


Figura 4.2: Niveles de los contaminantes NO2, NOX, PM10, y PM2.5 durante las primeras 6 semanas del 2017.

4.1.4 Implementación de modelos

Después de haber generado gráficos para la visualización de la evolución de las variables, se procede a generar modelos para el estudio de la relación entre las variables. Para ello se utiliza `Python` y la librería `Statsmodels` 4.1. Los tipos de modelos generados son los siguientes:

- Regresión lineal.
- Regresión lineal múltiple.
- ARIMA.

4.2 Implementación de la solución propuesta

En implementación de la solución propuesta se muestra el desarrollo realizado de los puntos planteados en 4.1. El desarrollo del presente proyecto se encuentra en el siguiente repositorio Github: `https://github.com/selenebpradop/relaciones-contaminantes-salud/`.

En la función 4.1 se muestra el proceso realizado para el procesamiento y agrupamiento de los datos en semanas epidemiológicas.

El fragmento de código 4.2 muestra como es que se generan las series de tiempo, esto después de haber procesado y agrupado los datos.

La función mostrada en 4.3 genera gráficos de radar al ingresarle como parámetros los datos ya procesados y agrupados.

En el fragmento de código 4.4 se muestra como son generados los modelos de regresión lineal después de haber procesado y agrupado los datos.

Código 4.1: Procesamiento y agrupamiento de datos

```
import pandas as pd
from epiweeks import Week, date
from sklearn import preprocessing
import seaborn as sns
import matplotlib.pyplot as plt
import string

columns = ['timestamp', contaminante]
dataframec = pd.read_csv('filled.csv', usecols=columns).dropna()
strfdt = '%d-%b-%y %H'
dataframec['timestamp'] = pd.to_datetime(dataframec['timestamp'], errors = 'coerce', format=strfdt)
dataframec = dataframec.dropna()
dataframec = dataframec.reset_index(drop=True)
dataframec['timestamp'] = dataframec['timestamp'].apply(lambda x: x.strftime('%Y-%m-%d %H'))
dataframeca = dataframec.loc[dataframec['timestamp'].str.startswith(año)]
dataframeca = dataframeca.reset_index(drop=True)
strfdt = '%Y-%m-%d %H'
dataframeca['timestamp'] = pd.to_datetime(dataframeca['timestamp'], errors = 'coerce', format=strfdt)
dataframeca['sem'] = dataframeca['timestamp'].apply(lambda x: date(x.year, x.month, x.day))
dataframeca['sem'] = dataframeca['sem'].apply(lambda x: Week.fromdate(x))
dataframeca['sem'] = dataframeca['sem'].apply(lambda x: x.week)
colums = ['EGRESO', 'DIAG_INI']
csvegresos = 'EGRESO_' + año + '.csv'
dataframeea = pd.read_csv(csvegresos, usecols=colums).dropna()
dataframeea['EGRESO'] = pd.to_datetime(dataframeea['EGRESO'], errors = 'coerce', format=strfdto)
dataframeea = dataframeea.loc[dataframeea['ENTIDAD'] == entidad]
dataframeea = dataframeea.dropna()
dataframeea = dataframeea.reset_index(drop=True)
numaño = int(año)
dataframeea['sem'] = dataframeea['EGRESO'].apply(lambda x: date(x.year, x.month, x.day))
dataframeea['sem'] = dataframeea['sem'].apply(lambda x: Week.fromdate(x))
dataframeea['sem'] = dataframeea['sem'].apply(lambda x: x.week)
dataframeea['EGRESO'] = dataframeea['EGRESO'].apply(lambda x: x if(x.year==numaño) else pd.NaT)
dataframeea = dataframeea.dropna()
dataframeea = dataframeea.reset_index(drop=True)
dataframesca = pd.DataFrame()
dataframesca['sem'] = semanas.index
dataframesca[contaminante] = ''
n = len(semanas.index)
for i in range (n):
    registrossem = dataframeca.loc[dataframeca['sem'] == i+1]
    promediocas = registrossem[contaminante].mean()
    dataframesca[contaminante][i] = promediocas
```

Código 4.2: Generación de series de tiempo

```
import pandas as pd
from epiweeks import Week, date
from sklearn import preprocessing
import seaborn as sns
import matplotlib.pyplot as plt
import string

diagnosticosaño = dataframeea['DIAG_INI'].value_counts()
diagnosticosaño = diagnosticosaño.sort_values(ascending = False)
ciesaño = dataframeea.groupby(['DIAG_INI', 'sem']).count()

s_scaler = preprocessing.StandardScaler()
ind = []
```

```
n = len(semanas.index)
for i in range (n):
    ind.append(i+1)
letras = []
for letra in string.ascii_uppercase:
    letras.append(str(letra))
# Se inicia un contador para controlar la cantidad de graficos a generar
cont = 0
maximo = 10
mindividuales = 7

# Proceso de generación de las figuras
print('\n' + año)
for name in diagnosticosaño.index:
    if cont < maximo:
        dataframegraficoacc = pd.DataFrame()
        dataframegraficoacc[contaminante] = dataframesca[contaminante]
        dataframegraficoacc = dataframegacc.reindex(ind)
        if cont < mindividuales:
            dataframegacc[name] = ciesaño['EGRESO'][name]
            for i in range (n):
                dataframegacc[contaminante][i+1] = dataframesca[contaminante][i]
            col_names = [contaminante, name]
        else:
            nameg =  letras[cont]
            ciesagrupadas = dataframeea.loc[dataframeea['DIAG_INI'].str.startswith(nameg)]
            ciesagrupadas = ciesagrupadas['sem'].value_counts()
            dataframegacc[nameg] = ciesagrupadas
            for i in range (n):
                dataframegacc[contaminante][i+1] = dataframesca[contaminante][i]
            col_names = [contaminante, nameg]
        df_s = s_scaler.fit_transform(dataframegacc)
        df_s = pd.DataFrame(df_s, columns=col_names)
        fig, ax = plt.subplots(ncols=1, figsize=(20, 8))
        print('\n' + col_names[0] + ' & ' + col_names[1])
        ax.set_title('Contaminante ' + col_names[0] + ' & CIE ' + col_names[1])
        ax.set_xlabel('Semana del año ' + año)
        sns.kdeplot(data=df_s)
        plt.savefig(contaminante + '/' + col_names[0] + '&' + col_names[1] + '_' + año + '.jpg', format='jpg')
        plt.show()
    cont = cont+1
```

Código 4.3: Generación de graficos de radar

```
def create_spiderwebs(datasets, lenlines, numspiders, title, titles, spoke_labels, colors, typeframe):

    # Set the number of lines of each spiderweb
    N = len(datasets)
    theta = radar_factory(N, frame=typeframe)
    # Set the number of columns and rows
    if (numspiders%2==0):
        numrows = 2
        numcols = int(numspiders/2)
    else:
        numrows = 1
        numcols = numspiders

    # Draw the shape of the spiderweb
    fig, axs = plt.subplots(figsize=(8, 8), nrows=numrows, ncols=numcols, subplot_kw=dict(projection='radar'))
    fig.subplots_adjust(wspace=0.5, hspace=0.20, top=0.85, bottom=0.05)
    newn = 0.5
```

```
    rgrids = []
    for z in range(lenlines*2):
        rgrids.append(newn)
        newn = newn + 0.5

    # Counter of the number of spiders
    i=0

    # Plot each case on separate axes
    for ax, (titlespiderweb) in zip(axs.flat, titles):
        # Put labels in the lines
        ax.set_rgrids(rgrids)
        ax.set_title(titlespiderweb, weight='bold', size='medium', position=(0.5, 1.1))
        dataspider = []
        # Normalize data
        for y in range(N):
            currentdata = datasets[y]
            number = currentdata[i]
            nmin = min(currentdata);
            nmax = max(currentdata);
            r = nmax - nmin
            x = (number-nmin)/r
            y = lenlines*x
            dataspider.append(y)
        # Draw the new lines in the spiderweb
        ax.plot(theta, dataspider, color=colors[i])
        ax.fill(theta, dataspider, facecolor=colors[i], alpha=0.25)
        # Put the name of each line in the figure
        ax.set_varlabels(spoke_labels)
        # Increment the counter
        i=i+1
    # Show the figure
    plt.show()
```

Código 4.4: Generación de los modelos de regresión lineal

```
datos = pd.DataFrame(dataframegacc, columns=col_names)
# Gráfico
# ==============================================================================
fig, ax = plt.subplots(figsize=(6, 3.84))
datos.plot(
    x    = col_names[0],
    y    = col_names[1],
    c    = 'firebrick',
    kind = "scatter",
    ax   = ax
)
ax.set_title('Contaminante ' + col_names[0] + ' & CIE ' + col_names[1])

# Correlación lineal entre las dos variables
# ==============================================================================
corr_test = pearsonr(x = datos[col_names[0]].fillna(0), y =  datos[col_names[1]].fillna(0))
print("Coeficiente de correlación de Pearson: ", corr_test[0])
print("P-value: ", corr_test[1])

# División de los datos en train y test
# ==============================================================================
xx = datos[[col_names[0]]].fillna(0)
yy = datos[[col_names[1]]].fillna(0)

x2 = xx.values.tolist()
```

```
y2 = yy.values.tolist()

# Creación del modelo utilizando matrices como en scikitlearn
# ==============================================================================
# A la matriz de predictores se le tiene que añadir una columna de 1s para el intercept del modelo
x = np.array(x2).astype(float)
y = np.array(y2).astype(float)
#ones = np.ones(len(x[0]))
#X = sm.add_constant(np.column_stack((x[0], ones)))
#for ele in x[1:]:
#    X = sm.add_constant(np.column_stack((ele, X)))
modelo = sm.OLS(y, x)
modelo = modelo.fit()
print(modelo.summary())
```

Capítulo 5

Experimentos

En el presente capítulo se presenta el diseño de los experimentos realizados así como los resultados obtenidos de ellos.

En esta sección se tratan los resultados obtenidos partiendo de desarrollar algunos experimentos que permiten determinar si la solución propuesta cumple con el objetivo planteado.

5.1 Diseño experimental

En la presente sección se discute el diseño de experimentos, es decir, que valores constantes fueron utilizados para su realización y por que se usan dichos valores.

5.1.1 Datos de entrada

Los datos de egresos hospitalarios del año 2014 al año 2018 provienen de la base de datos de la Secretaría de Salud del Gobierno de México [9]. También se tienen registros de los niveles de PM10, PM2.5, NOX Y NO2 presentes en el área metropolitana de Monterrey, dichos registros son hechos por las estaciones de mo-

nitoreo pertenecientes al SIMA [20] mostradas en la figura 1.1. Los documentos con los datos son proporcionados por Benavides.

Selección de datos. Los conjuntos de datos por año de egresos hospitalarios contienen información de todos los estados de México, por lo cual se hace una limpieza de datos para solo obtener los registros de Nuevo León ya que de dicha entidad es de la cual se tienen los datos de contaminación.

Datos de ingresos hospitalarios. Se agrupan en semanas epidemiológicas para una mejor manipulación de ellos. Por CIE [18] se obtiene el numero de egresos en cada semana del año.

Datos de los contaminantes. Se agrupan en semanas epidemiológicas para una mejor manipulación de ellos. Se obtiene el promedio del nivel del contaminante por cada semana del año.

5.1.2 Visualización de datos

Con los datos ya seleccionados y agrupados se procede a generar las visualizaciones de los datos. Las visualizaciones de los datos se hacen con series de tiempo y gráficos de radar.

Series de tiempo. Se procede a generar series de tiempo de cada año por contaminante y CIE [18]. Las variables ajustables son:

- contaminante: El nombre del contaminante.
- año: El año del que se quieren obtener las series de tiempo.
- numcies: Número de series de tiempo a generar por contaminante. Se parte de la CIE [18] con mayor numero de egresos.

Gráficos de radar. Los datos ya seleccionados y agrupados se normalizan teniendo como valor mínimo cero y como valor máximo un numero entre uno y cuatro. Posteriormente se generan gráficos de radar de cada año por semana, en las que se muestra el nivel contaminante y la variación de las CIES [18]. Las variables ajustables son:

- Número de series de tiempo a generar por contaminante. Se parte de la CIE [18] con mayor numero de egresos.
- Valor máximo que se utiliza para representar la longitud de los ejes en el gráfico.
- Nombre de la figura.
- Nombre de cada eje en el gráfico.
- Los colores de cada eje en el gráfico.
- Si el spiderweb es generado de forma circular o en forma de polígono.

5.1.3 Generación de modelos

Se procede a generar los modelos. En cada modelo se tienen métricas para evaluar su eficacia y valores que pueden ser ajustados en función de encontrar la combinación que proporcione mejores resultados.

5.1.3.1 Regresión lineal

Se tienen algunas variables que pueden ser modificadas para la generación de los modelos de regresión lineal.

- Número de series de tiempo a generar por contaminante. Se parte de la CIE [18] con mayor numero de egresos.

- Porcentaje de datos utilizados para el entrenamiento del modelo.
- Nivel de significancia.

5.2 Resultados

Establecidos los experimentos que se van a realizar, se reporta los resultados obtenidos...

5.3 Discusión

Todos los experimentos son ejecutados en una laptop con las especificaciones del cuadro 5.1.

Cuadro 5.1: Especificaciones técnicas del equipo de cómputo

Sistema Operativo	macOS Big Sur
Procesador	Apple M1
RAM	8 GB RAM

Capítulo 6

Conclusiones

Este capítulo describe la tesis a partir de la manera que cumple los objetivos generales y específicos para determinar si la hipótesis se comprueba, trata también del porque se realizó la tesis...

6.1 Contribuciones

La solución propuesta surgió a partir de...

6.2 Trabajo a futuro

La solución propuesta en la tesis...

Bibliografía

[1] Arias, J. R. (2006), «What is an epidemiological week and why do we use them», *Skeeter*, **66**(1), pág. 7.

[2] Ballester Díez, F., J. M. Tenías y S. Pérez-Hoyos (1999), «Efectos de la contaminación atmosférica sobre la salud: una introducción», *Revista Española de Salud Pública*, **73**(2), págs. 109–121.

[3] Benavides, A. (2022), «Perfil de Github», `https://github.com/jbenavidesv87`.

[4] Breton, R. M. C., J. C. Breton, M. de la Luz Espinosa Fuentes, J. Kahl, A. A. E. Guzman, R. G. Martínez, C. Guarnaccia, R. del Carmen Lara Severino, E. R. Lara y A. B. Francavilla (2021), «Short-Term Associations between Morbidity and Air Pollution in Metropolitan Area of Monterrey, Mexico», *Atmosphere*, **12**(10), pág. 1352.

[5] Brockwell, P. J., P. J. Brockwell, R. A. Davis y R. A. Davis (2002), *Introduction to time series and forecasting*, Springer.

[6] Catalá, F. y E. De Manuel (1998), «Informe SESPAS 1998: La salud pública y el futuro del estado del bienestar», *Granada: EASP*.

[7] Corbitt, R. y R. Corbitt (1990), *Standard Handbook of Environmental Engineering*, McGraw-Hill.

[8] Darlington, R. B. y A. F. Hayes (2016), *Regression analysis and linear models: Concepts, applications, and implementation*, Guilford Publications.

[9] Dirección General de Información en Salud (2021), «Egresos Hospitalarios», URL http://www.dgis.salud.gob.mx/contenidos/basesdedatos/da_egresoshosp_gobmx.html.

[10] FC Lichtenfels, A. J., D. A. Van Der Plaat, K. de Jong, C. C. van Diemen, D. S. Postma, I. Nedeljkovic, C. M. van Duijn, N. Amin, S. la Bastide-Van Gemert, M. De Vries *et al.* (2018), «Long-term air pollution exposure, genome-wide DNA methylation and lung function in the LifeLines cohort study», *Environmental health perspectives*, **126**(2), pág. 027 004.

[11] Guarnaccia, C., J. G. C. Breton, R. M. C. Breton, C. Tepedino, J. Quartieri y N. E. Mastorakis (2018), «ARIMA models application to air pollution data in Monterrey, Mexico», en *AIP Conference Proceedings*, tomo 1, AIP Publishing LLC, pág. 020041.

[12] Gupta, A., H. Bherwani, S. Gautam, S. Anjum, K. Musugu, N. Kumar, A. Anshul y R. Kumar (2021), «Air pollution aggravating COVID-19 lethality? Exploration in Asian cities using statistical models», *Environment, Development and Sustainability*, **23**(4), págs. 6408–6417.

[13] Haddad, R. (1974), «Contaminación del aire. Situación actual en la América Latina y el Caribe», *Informe técnico.*

[14] Kim, J. S., Z. Chen, T. L. Alderete, C. Toledo-Corral, F. Lurmann, K. Berhane y F. D. Gilliland (2019), «Associations of air pollution, obesity and cardiometabolic health in young adults: The Meta-AIR study», *Environment international*, **133**, pág. 105 180.

[15] Korc, M. y R. Sáenz (1999), «Monitoreo de la calidad del aire en América Latina», *Korc Marcelo E*, págs. 1–22.

[16] Liu, Y., J. Pan, H. Zhang, C. Shi, G. Li, Z. Peng, J. Ma, Y. Zhou y L. Zhang (2019), «Short-term exposure to ambient air pollution and asthma

mortality», *American journal of respiratory and critical care medicine*, **200**(1), págs. 24–32.

[17] Martín, R. M. y M. S. Bayle (2018), «Impacto de la contaminación ambiental en las consultas pediátricas de Atención Primaria: estudio ecológico», en *Anales de Pediatría*, tomo 2, Elsevier, págs. 80–85.

[18] Organization, W. H. *et al.* (2016), «International Classification of Diseases. 2016», *World Health Organization*.

[19] Peniche-Camps, S. y M. Cortez-Huerta (2020), «La costumbre al envenenamiento: El caso de los contaminantes atmosféricos de la ciudad de Guadalajara, México», *Revista de Ciencias Ambientales*, **54**(2), págs. 1–19.

[20] SIMA (2015), «Sistema Integral de Monitoreo Ambiental», URL `http://aire.nl.gob.mx`.

[21] To, T., J. Zhu, D. Stieb, N. Gray, I. Fong, L. Pinault, M. Jerrett, A. Robichaud, R. Ménard, A. van Donkelaar *et al.* (2020), «Early life exposure to air pollution and incidence of childhood asthma, allergic rhinitis and eczema», *European Respiratory Journal*, **55**(2).

[22] Zhang, Z., B. Dong, S. Li, G. Chen, Z. Yang, Y. Dong, Z. Wang, J. Ma y Y. Guo (2019), «Exposure to ambient particulate matter air pollution, blood pressure and hypertension in children and adolescents: a national cross-sectional study in China», *Environment international*, **128**, págs. 103–108.

Resumen autobiográfico

Selene Berenice Prado Prado

Candidato para obtener el grado de
Ingeniería en Tecnología de Software

Universidad Autónoma de Nuevo León
Facultad de Ingeniería Mecánica y Eléctrica

Tesis:
Modelado y visualización de relaciones entre contaminantes del aire y salud pública

Nací el 30 de Junio de 2000 en Monterrey, Nuevo León, soy la mayor de cuatro hijos. Mi familia está conformada por mi madre Lilia Prado López, mi padre Adan Alfaro Lerma, y mis hermanos: Angel Alejandro Prado Prado, Estrella Belen Prado Prado, y Genesis Adali Alfaro Prado.

Desde pequeña me han gustado las matemáticas, aprender como funcionan los sistemas computacionales, y leer.

Durante los primeros semestres de mi carrera descubrí la inteligencia computacional, un área que me encantó desde que la descubrí, en especial su rama de ciencia de datos, rama en la que espero seguir desarrollándome.

Otra cosa que me apasiona es dibujar y pintar, actividades que estaban dentro de mi pero que se avivaron cuando inició la pandemia en el año 2020.